Prediction of Europe Football Outcomes using Deep Networks

Kornraphop Kawintiranon, Hanjing Wang, Armaan Khullar, Jiajia Liu
ANLY 590 - Neural Nets and Deep Learning | Georgetown University

Abstract

For most soccer statistics, the observations are very limited to aggregated data such as goals, shots, and fouls. However, a football game could generate much more events, and it is very important and interesting to take into account the context in which those events were generated. Our work involves three models. The ANN achieves approximately 66% of accuracy on our validation set for football outcome prediction. The CNN utilizes only textual features, and its performance using text alone is approximately 66% of accuracy on the validation set for outcome prediction. The RNN is established with an accuracy of 96% for predicting the event sequences. Finally, we designed an ensemble ANN model in which we combine the results from separate models to achieve a better performance. We aim to investigate outcomes of Europe Football teams and improve the prediction results.

Introduction

We propose methodologies and empirical results to better help with these predictions. We limit our study to European soccer and to the five largest European soccer leagues: England, Spain, Germany, Italy, and France, between 2011 and 2017. In our experiments, we implemented deep learning models to analyze both team-based and game-based data. During our experimentation, we found some interesting questions that we sought to answer:

1: What are the top teams based on their goals of performance?
2: Can we predict the events the sequence of events that would transpire during a game's second half?
3: Can we use commentary during a game's first half to predict the final outcome of the game?

We will first discuss the related work before going into the details of our feature engineering techniques, experiments, and results.

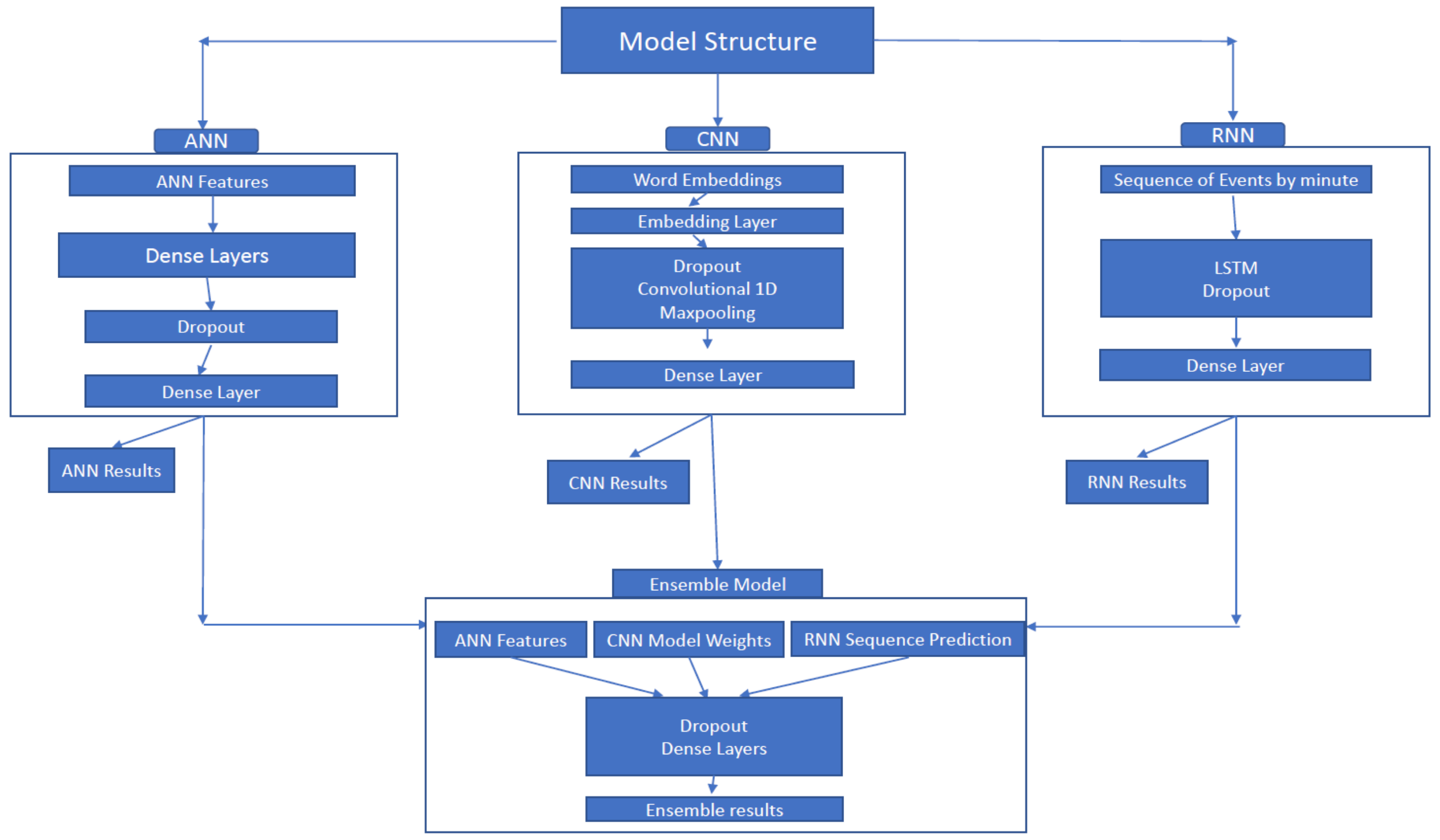
Dataset

For this project, we are focusing on the "events.csv" Football Events dataset that was provided by Kaggle. This dataset contains event data for 9,074 games for the top five European soccer leagues: England, Spain, Germany, Italy, and France, between 2011 and 2017. It records information pertaining to 16 types of events that transpired over the course of a game. For example, shot attempted, corner kick, foul, yellow card, second yellow card, red card given, substitution, free kick won, offsides, handball, penalty conceded, etc.

Related Work

The majority of research on this task has been done on behalf of researchers at private organizations or by freelancers. Publicly available work in this area is, therefore, limited. However, we found one paper called *Predicting Soccer Match Results in the English Premier League*, which was written by students at Stanford. In their research, the authors focused on soccer matches in the English Premier League. They use a variety of machine learning models. Nonetheless, we borrowed some ideas such as constructing a three-class classification problem for some of our own models. Finally, we were also influenced by the work Matheus Gonzaga and ALexandra Rodriguez from their blog *We used Neural Networks to predict the 2018 World Cup champion*, on the website Poatek. In particular, they used a Recurrent Neural Network in order to predict the winner of the 2018 World Cup.

Methodology



Artificial Neural Networks (Model 1):

We constructed was a three-classification problem in which we predict which team wins or tie. This involved feature engineering of the original data in which we create both game-based and team-based features. We first generate 38 game-based features The team-based data with 10 features are generated in which the average goals and the winning percentage for each team are also considered. The model has a total of 48 features. After feature preparation, an ANN model is established. The model has several dense layers with 64 nodes and a rectifier activation function for each layer. We also added dropout regularization to reduce the overfitting problem.

Final Ensemble Model: Combining all models

We combine the separate models by adding the new features generated in CNN and RNN model to the original ANN three-classification model. In particular, we use the CNN model's weights and event sequence prediction results from the LSTM as the additional features. We thus have 343 features in the final ANN model.

Convolutional Neural Networks (Model 2):

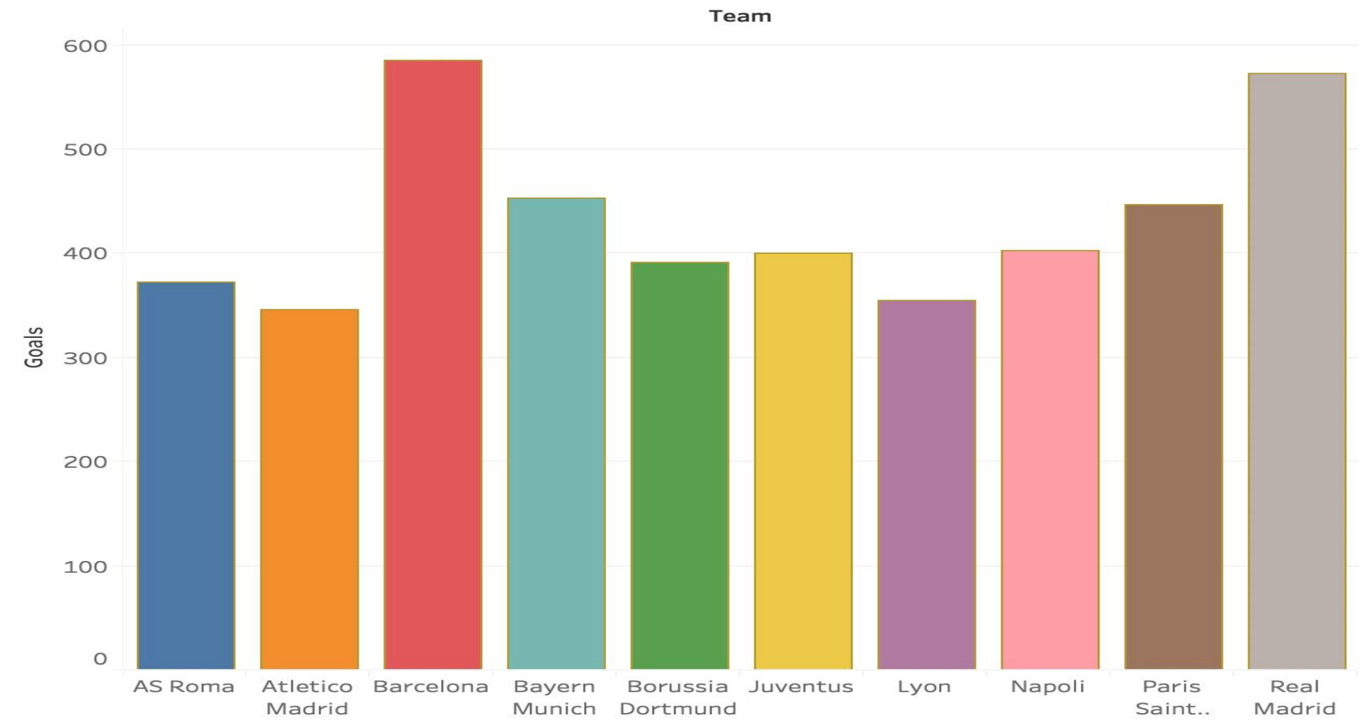
In this model, we aggregate the comments from commentators of all events that occurred during the first half of each game in order to predict the team that will win the game. We build word embeddings based on all the game's comments and feed them into the CNN model. The result of each game is a probability of the home team winning, away team winning, or the game ending in a tie.

Recurrent Neural Networks (Model 3):

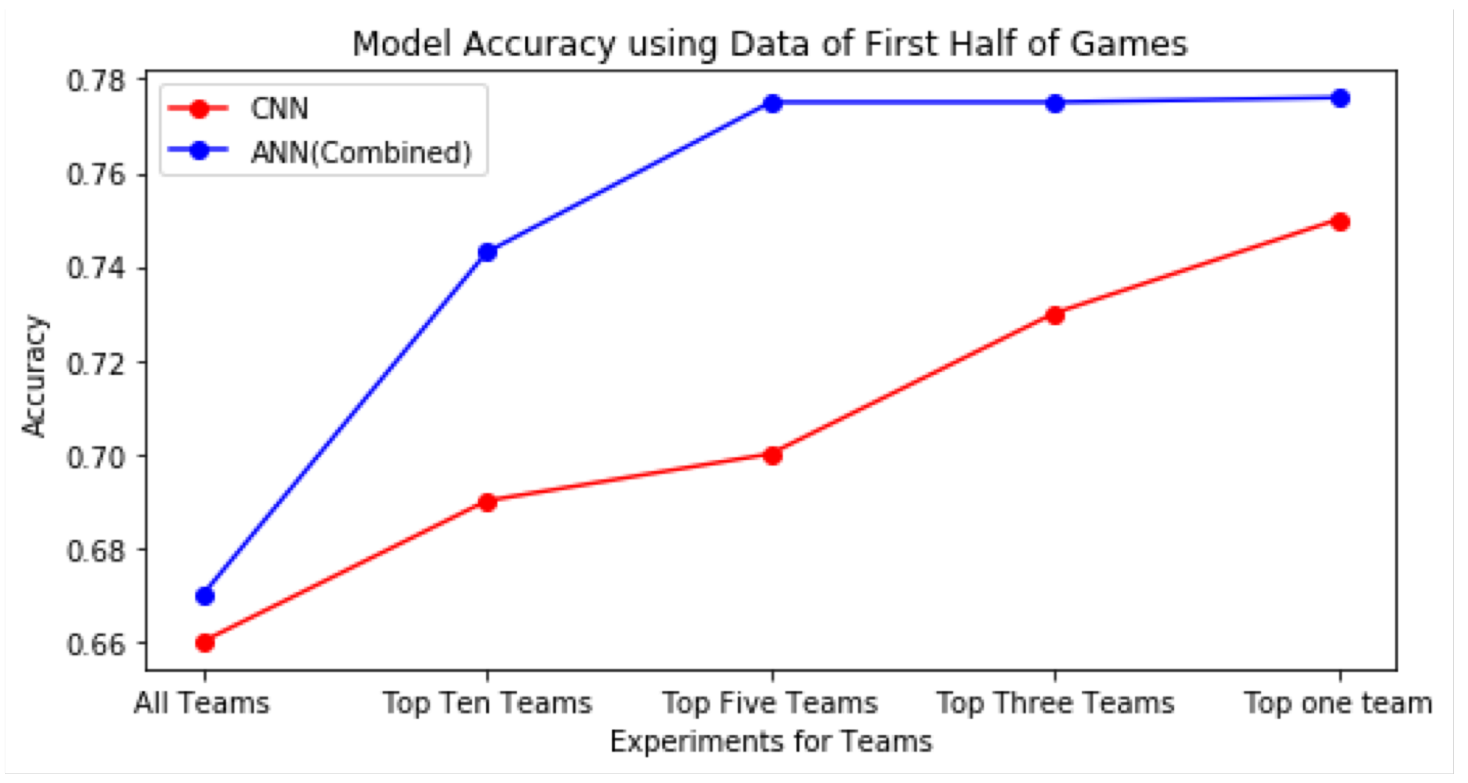
In this model, we try to predict the sequence of events that will transpire during the second half of a soccer game using information from the first half. In this experiment, we create a Long short-term memory (LSTM) model to predict a game's sequence of events for each minute of the second half using the sequence of events from the first half. We initially try to predict whether each of the events occur for each minute of a game, regardless of the team involved. We then double our original event features by creating a separate event attribute dedicated to each team. We also add in other features such as "assist method" and "fast break". We train our model and predict the sequence of events for each minute of the second half for each game. .

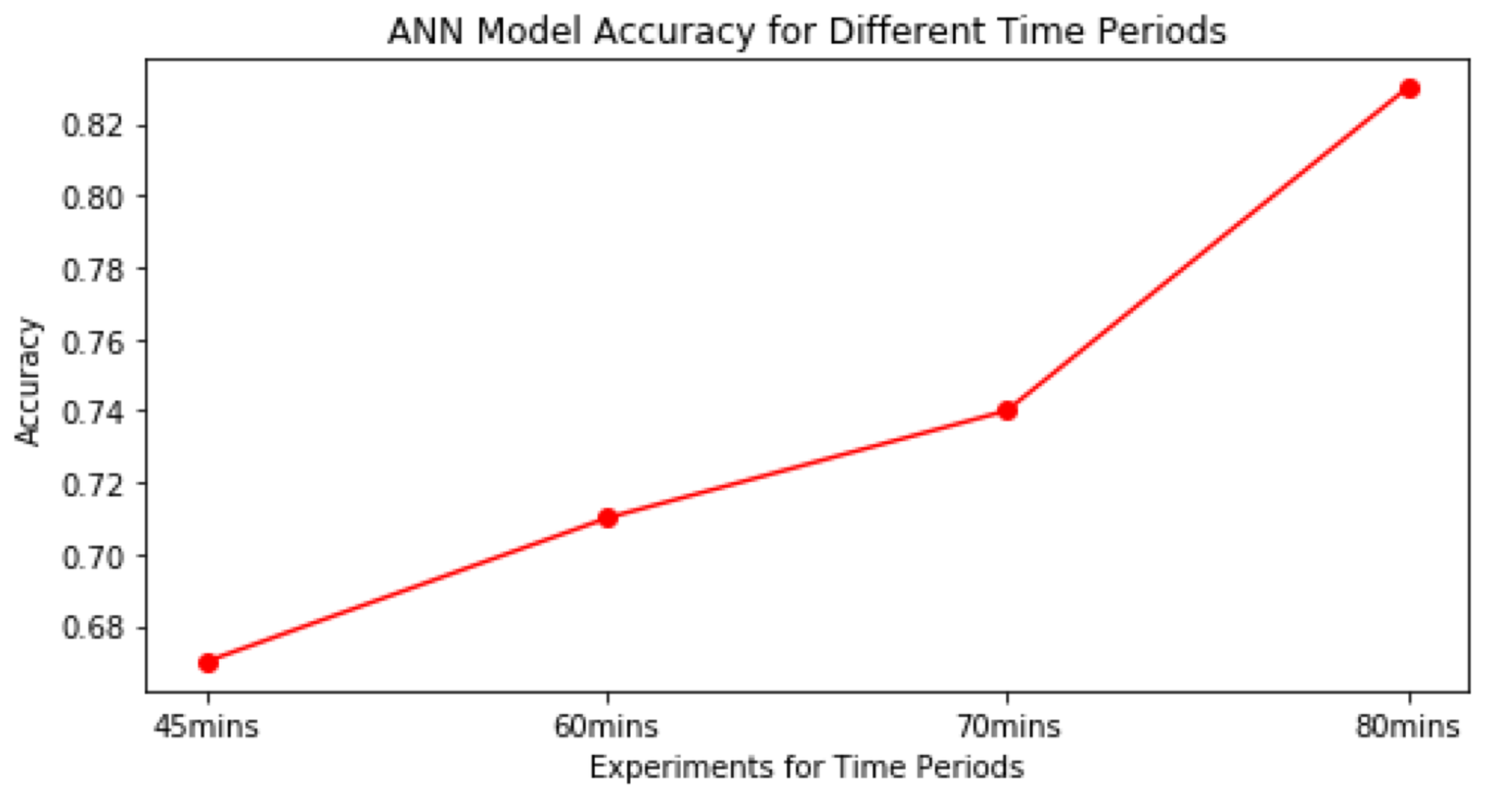
Results

Top 10 Team's Performances



Outcome Prediction





Conclusion

Overall, it seems that we have very interesting results and observations. We utilize textual comments from commentators during the first half of each game, in order to predict the outcome of a game (win-loss-tie). First, we use only the textual data with the CNN to perform the predictions. The model accuracy is only 66% with a randomized validation set. Furthermore, preliminary results suggest that the ANN model performs better if we consider more minutes per game. When we combine the game-based and team-based features of the ANN, with the predictions generated by the CNN and the LSTM, we create our ensemble model. We see that it performs better if we consider only the top teams.